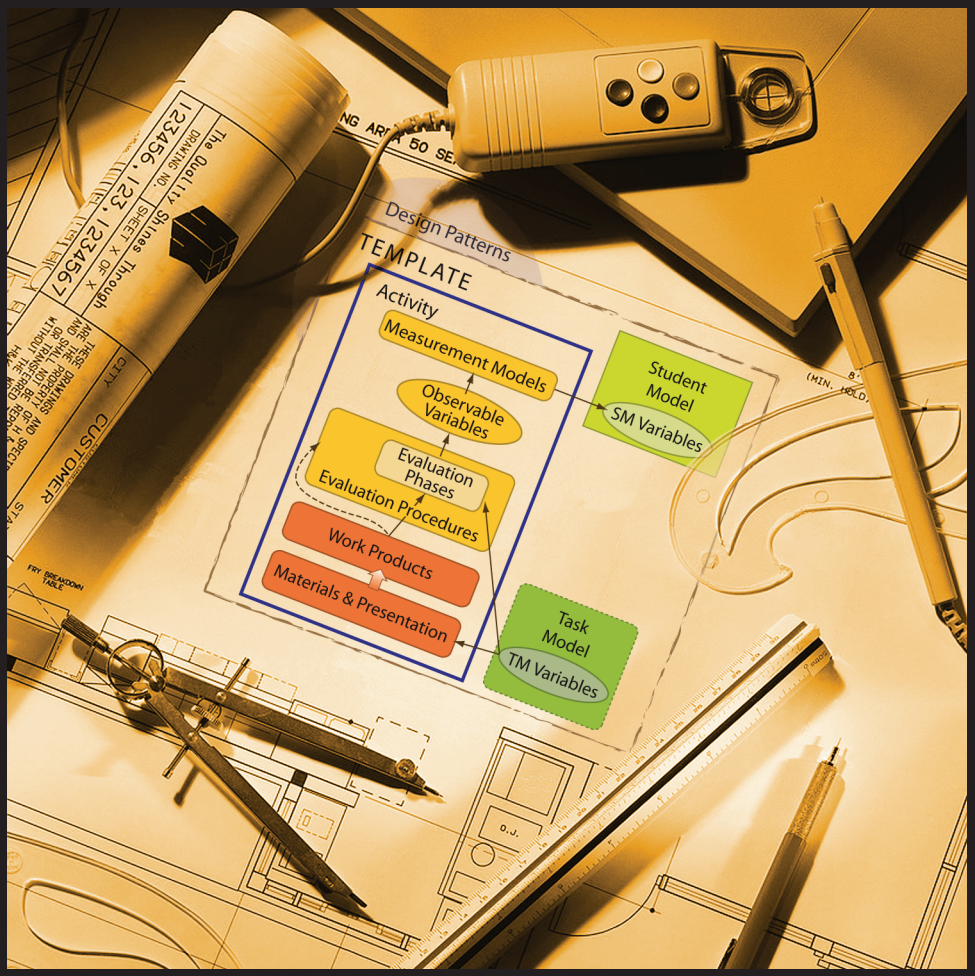


PADI Technical Report 17 | September 2006

Implications of Evidence-Centered Design for Educational Testing

PADI | Principled Assessment Designs for Inquiry

Robert J. Mislevy, University of Maryland

Geneva D. Haertel, SRI International

Report Series Published by SRI International

SRI International
Center for Technology in Learning
333 Ravenswood Avenue
Menlo Park, CA 94025-3493
650.859.2000
http://padi.sri.com

PADI Technical Report Series Editors
Alexis Mitman Colker, Ph.D., *Project Consultant*
Geneva D. Haertel, Ph.D., *Co-Principal Investigator*
Robert Mislevy, Ph.D., *Co-Principal Investigator*
Klaus Krause, *Technical Writer/Editor*
Lynne Peck Theis, *Documentation Designer*

PRINCIPLED ASSESSMENT DESIGNS FOR INQUIRY
TECHNICAL REPORT 17

Implications of Evidence-Centered Design for Educational Testing

DRAFT

Prepared by:
Robert J. Mislevy, University of Maryland
Geneva D. Haertel, SRI International

Acknowledgments
This material is based on work supported by the National Science Foundation under grant REC-0129331 (PADI Implementation Grant).
Disclaimer
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

FIGURES

TABLES

ABSTRACT

Evidence-centered assessment design (ECD) provides language, concepts, and knowledge representations for designing and delivering educational assessments, all organized around the evidentiary argument an assessment is meant to embody. This article describes ECD in terms of layers for analyzing domains, laying out arguments, creating schemas for operational elements such as tasks and measurement models, implementing the assessment, and carrying out the operational processes. It is argued that this framework helps designers take advantage of developments from measurement, technology, cognitive psychology, and learning in the domains. Examples of ECD tools and applications are drawn from the Principled Assessment Designs for Inquiry (PADI) Project. Attention is given to implications for large-scale tests such as state accountability measures, with a special eye for computer-based simulation tasks.

1.0 Introduction

These are heady times in the world of educational assessment—days of urgent demands, unprecedented opportunities, and tantalizing challenges. The demands are for consequential tests in schools and states, at larger scales and with higher stakes than we have seen before. The opportunities are to assess learning viewed from a growing understanding of the nature and acquisition of knowledge and to draw upon ever-expanding technological capabilities to construct scenarios, interact with examinees, capture and evaluate their performances, and model the patterns they convey. And the challenges are abundant, encapsulated in a single question: How can we bring these new capabilities to bear on the assessment problems we face today?

Long established and well-honed assessment practices did not evolve to deal with assessments that are complex, in the sense of interactive tasks, multidimensional proficiencies, and complex responses to evaluate. But progress is being made on many fronts to extend practice, as seen in the National Board of Examiners' computer-based simulation tasks (Clyman, Melnick, & Clauser, 1999), Adams, Wilson, and Wang's (1997) structured multidimensional IRT models, and White and Frederiksen's (1998) guided self-evaluation in extended inquiry tasks. This work succeeds because even when it differs from traditional testing on the surface, each innovation is grounded in the same principles of evidentiary reasoning that underlie the best assessments of the past.

One vital line of current research aims to make the principles of evidentiary reasoning explicit, to build conceptual and technological tools that help designers orchestrate new developments, and lay the groundwork for further advances. The National Research Council's (2001) volume *Knowing What Students Know* lays out the case and provides an integrative review of the necessary cognitive, psychometric, and technological foundations. Examples of work that coordinate various aspects of task design, psychometric modeling, assessment delivery, and psychological research in the desired ways include Baker (1997, 2002), Embretson (1985, 1998), Luecht (2002), and Wilson (2005).

Our own recent work falls under the rubric of "evidence-centered" assessment design (ECD; Mislevy, Steinberg, & Almond, 2003), an approach that has been implemented in various ways at Educational Testing Service, Cisco Systems (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004), the IMS Global Learning Consortium (2000), and elsewhere. We will illustrate points with tools and examples from our ECD-based work in the Principled Assessment Designs for Inquiry (PADI; Baxter & Mislevy, 2004) project. The BioKIDS project's (Songer, 2004) application of PADI design tools illustrates their benefits in large-scale on-demand testing, a particular focus of this presentation.

The next section of the paper provides a brief overview of evidence-centered design. Two complementary ideas organize the effort. The first is an overarching conception of assessment as an argument from imperfect evidence. Messick (1994) lays out the basic narrative, saying that we:

> would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or

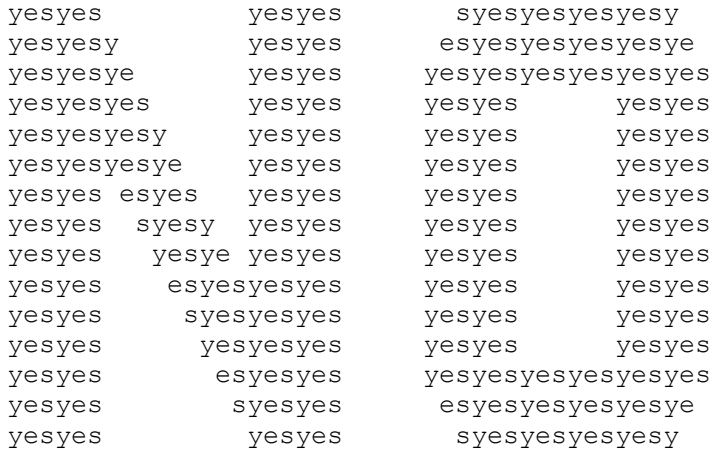
performances should reveal those constructs, and what tasks or situations should elicit those behaviors? (p. 16).

The second idea is distinguishing layers at which activities and structures appear in the assessment enterprise, all for the purpose of instantiating an assessment argument in operational processes (Mislevy, Steinberg, & Almond, 2003; Mislevy & Riconscente, 2006).

The following section steps through the layers of evidence-centered design as applied to assessment in more detail. We see the roles that advances from allied fields bring to assessment and how their contributions are coordinated within and across layers. The benefits of explicitness, reusability, and common language and representations are noted throughout.

The closing discussion addresses a question posed by an anonymous reviewer of a symposium proposal that we submitted for an annual meeting of the National Council of Measurement in Education (NCME): Isn't this all just new words for what people are already doing? To anticipate, the answer is symbolized in Figure 1: Many small "yes's," arranged in just the right way, reveal a greater "no."

Figure 1. "Yes" Means "No"



1.1 Evidence-Centered Assessment Design

Evidence-centered design (ECD) views an assessment as an evidentiary argument: An argument from what we observe students say, do, or make in a few particular circumstances, to inferences about what they know, can do, or have accomplished more generally (Mislevy, Steinberg, & Almond, 2003). The view of assessment as argument is a cornerstone of test validation (Kane, 1992, Messick, 1989). ECD applies this perspective proactively to test design.

2.0 Layers in Assessment Design

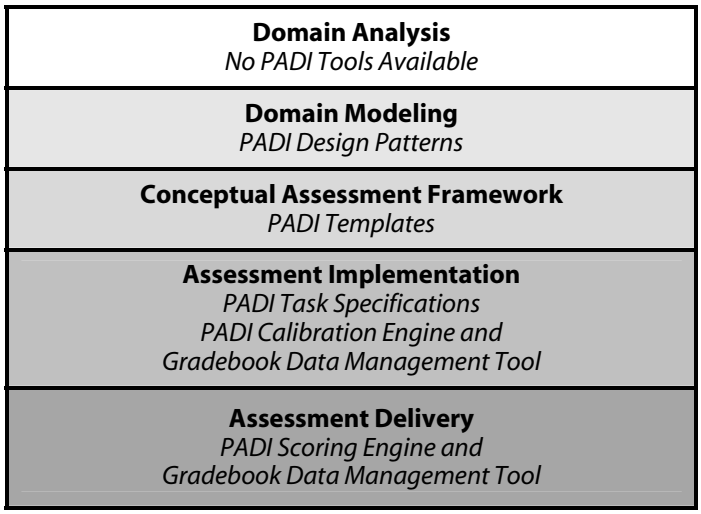
ECD organizes the work of design and implementation in terms of layers, a metaphor drawn from architecture and software engineering (Mislevy & Riconscente, 2006). It is often useful to analyze complex systems in terms of subsystems whose individual components are better handled at the subsystem level (Simon, 2001). Brand (1994) views buildings as dynamic objects wherein initial construction and subsequent changes take place at different timescales and in varying ways, by actors with different motives and roles. Brand identifies six layers that capture the stages of design and implementation of a building. These layers serve as a heuristic for making decisions at each step in the life of a building. To support maintenance and troubleshooting, Cisco System's (2000) Open System Interconnection (OSI) reference model distinguishes seven layers of activity in computer networks: physical, data link, network, transport, session, presentation, and application. Tasks are self-contained within each, so network functions within each layer can be implemented independently and updated without impacting the other layers. In both cases, certain processes and constraints are in place within each layer while cross-layer communication is limited and tuned to the demands of the overall goal. Knowledge representations, work flow, and communications are organized in terms of the layers.

Evidence-centered design applies the concept of layers to rationalize the complex process of designing, implementing, and delivering an educational assessment. ECD identifies five layers. Each is characterized in terms of its role in the assessment development process, the key concepts, tools, and entities used at each layer, and common knowledge representations that assist in achieving each layer's purpose. The layers are *domain analysis*, *domain modeling*, *conceptual assessment framework, assessment implementation*, and *assessment delivery*. Although the layering suggests a sequential design process, cycles of iteration and refinement both within and across layers are typical. Table 1 summarizes the ECD layers and their roles, key entities, and examples of knowledge representations. Figure 2 illustrates the relationship among the layers and notes PADI tools or resources that are available to support design activities in each layer.

Table 1. Layers of Evidence Centered Design for Educational Assessments

Layer	Role	Key Entities	Selected Knowledge Representations
Domain Analysis	Gather substantive information about the domain of interest that has direct implications for assessment; how knowledge is constructed, acquired, used, and communicated.	Domain concepts; terminology; tools; knowledge representations; analyses; situations of use; patterns of interaction.	Representational forms and symbol systems used in domain (e.g., algebraic notation, Punnet squares, maps, computer program interfaces, content standards, concept maps).
Domain Modeling	Express assessment argument in narrative form based on information from domain analysis.	Knowledge, skills and abilities; characteristic and variable task features, potential work products, potential observations.	Toulmin and Wigmore diagrams; PADI design patterns; assessment argument diagrams; "big ideas" of science
Conceptual Assessment Framework	Express assessment argument in structures and specifications for tasks and tests, evaluation procedures, measurement models.	Student, evidence, and task models; student, observable, and task variables; rubrics; measurement models; test assembly specifications; PADI templates and task specifications	Algebraic and graphical representations of measurement models; PADI task template; item generation models; generic rubrics; algorithms for automated scoring.
Assessment Implementation	Implement assessment, including presentation-ready tasks and calibrated measurement models.	Task materials (including all materials, tools, affordances); pilot test data to hone evaluation procedures and fit measurement models.	Coded algorithms for rendering tasks, interacting with examinees & evaluating work products; tasks as displayed; IMS/QTI representation of materials; ASCII files of item parameters.
Assessment Delivery	Coordinate interactions of students and tasks: task-and test-level scoring; reporting	Tasks as presented; work products as created; scores as evaluated.	Renderings of materials; numerical and graphical summaries for individual and groups; IMS/QTI results files

Figure 2. Evidence-Centered Design Layers and Associated PADI Tools



2.1 Domain Analysis

The *domain analysis* layer requires gathering substantive information about the domain that is to be assessed. If the assessment being designed is to measure science inquiry at the middle school level, the *domain analysis* activity would focus on the gathering of information about the concepts, terminology, representational forms, and ways of interacting that professionals working in the domain use and that educators have found useful to help students acquire this knowledge.

Examples of *domain analysis* can be found in the work of Webb (2006) who has described the content to be assessed in measures of achievement testing. Documents such as the National Science Education Standards (National Research Council, 1996) often provide a good starting point (state standards documents are in fact mandated foundations of accountability tests). In the area of language testing, the Bachman and Palmer (1996) taxonomy of task characteristics can be used at both the *domain analysis* and the *conceptual assessment framework* levels. These language testing experts use the taxonomy to describe both the features of target language use (TLU) and the intended assessment, and they establish a correspondence between the two. *Domain analysis* has also been compared to aspects of practice analysis for credential testing (Raymond & Neustel, 2006). Practice analysis uses rich task descriptions for the purpose of extracting features of tasks that are required for successful completion of certain jobs. The task features help assessment designers identify the kinds of Knowledge, Skills, and Abilities (KSAs) they will need to draw inferences about in order to serve the purposes of the assessment. Automated methods for carrying out *domain analysis*, such as Shute, Torreano, and Willis's (2000) automated knowledge elicitation tool DNA (for Decompose, Network, Assess) can be tuned to provide input for *domain modeling*.

In addition to articulating the content as part of the *domain analysis*, it is also important to specify the psychological perspective assumed in instruction and assessment. A lack of alignment in the psychological perspective at different layers of the design process erodes

the coherence of the assessment argument for grounding the claims that the designer wants to make about examinees. Mislevy and Riconscente (2006) warn that "the psychological perspective greatly influences the overall assessment process and cannot be emphasized strongly enough" (p. 68). Being aware of whether the assessment is based on a behavioral, information processing, or sociocultural perspectives is crucial. In mathematics, a behaviorist perspective would lead to an assessment that found evidence that students could solve classes of mathematics problems by assembling stimulus-response bonds—memorizing and applying algorithms. An information processing approach would emphasize knowledge structures that underlie mathematics and the representational forms and cognitive processes by which students acquire and use them to solve problems (e.g., VanLehn, 1990). Assessment designers with an information processing perspective might look for evidence of reasoning patterns that lead to the desired understandings rather than common misconceptions. An assessment based on a sociocultural perspective would look still further to mathematics as it functions within a community of practice and fluency with the protocols and the above-mentioned forms as they are used in that setting (e.g., Lehrer & Schauble, 2002).

Transdisciplinary research on learning from the information-processing and sociocultural perspectives tells us much about how students become proficient in particular domains and, thus, what we need to assess. As the American Association for the Advancement of Science (1993) put it:

> Some powerful ideas often used by mathematicians, scientists, and engineers are not the intellectual property of any one field or discipline. Indeed, notions of system, scale, change and constancy, and models have important applications in business and finance, education, law, government and politics, and other domains, as well as in mathematics, science, and technology. These common themes are really ways of thinking rather than theories or discoveries. (p. 261).

PADI's applications in science revolve around paradigmatic ways of thinking, such as inquiry cycles (White & Frederiksen, 1998), knowledge representation (Markman, 1999; Greeno, 1983), model-based reasoning (Stewart & Hafner, 1994), and scaffolded learning (Brown, Collins, & Duguid, 1989). The BioKIDS project, for example, helps students learn to design investigations through the use of increasingly independent investigations (Huber, Songer, & Lee, 2003). Consequently, the assessment tasks BioKIDS builds probe the degree of support that students need to, say, build scientific explanations. We see in the next section how *design patterns* can leverage these recurring themes for building assessment tasks in different domains, for different purposes, and at different educational levels.

Work at the *domain analysis* layer holds important implications for assessment in identifying and synthesizing developments from many directions that bear on learning and therefore on assessment. Indeed, the same research and insights ought to be informing curriculum, instruction, and assessment in coordination, revolving around the same views of knowledge, how it develops, how it is used, and how it is manifest. For assessment specifically, there is often a great deal of information available about a domain and learning in the domain, but it is not organized along the lines of assessment arguments. This grounding is relevant to assessments of all kinds, whether formative or

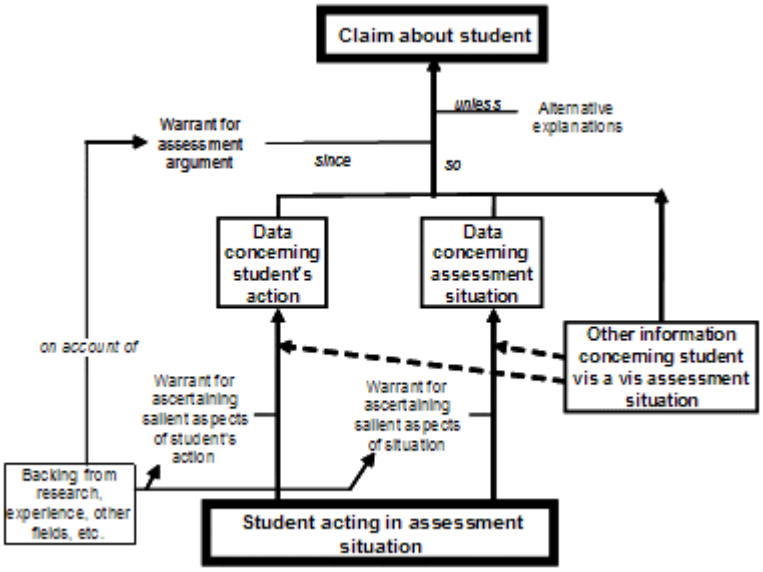
summative, large-scale or classroom, elementary students or candidates for advanced certifications. As the first stage in assessment design, *domain analysis* marshals the information that will provide the grounding for assessment designs. It leads us to understand what knowledge structures are in use in a domain, what is valued knowledge and work, task features, common representational forms, and performance outcomes. These kinds of information represent what is valued in this domain by teachers, researchers, and domain experts. As we move through the assessment design process, these categories of information presage the entities and structures that appear in the *domain modeling* design layer.

2.2 Domain Modeling

In the *domain modeling* layer we organize the information and relationships discovered in *domain analysis* into the shape of assessment arguments. There is a transition from the substantive, specialized compendium of knowledge about the target domain to forms that guide the building of the assessment machinery in the layer labeled the *conceptual assessment framework*. Tools that can be used to help carry out *domain modeling* are *diagrams of assessment arguments* (Kane, 1992), assessment argument schemas based on the "big ideas" of a given domain (Chung, Delacruz, Dionne, & Bewley., 2003), and *design patterns*, an approach developed in the PADI Project that lays out, in narrative form, key elements in an assessment argument.

Toulmin (1958) sets forth general structures for arguments in terms of claims, data, and warrants. Using Toulmin's terminology, Figure 3 shows the structure of an assessment argument. In assessment design, claims are the target of the assessment, such as a level of proficiency in solving quadratic equations or designing an experiment in genetics. The data are the salient aspects of the settings students act in (including goals, materials, and resources) and students' performances in those settings. The warrant is the logic or reasoning that explains why particular data provide appropriate evidence for the claims. We see at this point the involvement of both a view of the nature of knowledge and how it is used and the particular knowledge and capabilities in the target domain. As a formal structure, Figure 3 has connections to both the layers above it and below it. Looking back to *domain analysis*, it provides slots for various kinds of information from the domain in terms of their role in the assessment argument. Looking forward to the *conceptual assessment framework* (CAF), these roles will be instantiated by various pieces of assessment machinery.

Figure 3. An Extended Toulmin Diagram for Assessment Arguments



As assessment diagrams provide support for understanding the structure of an assessment argument, *design patterns* provide support for its substance. The preceding section noted the common themes in reasoning in science and other domains. Similarly, expertise research has provided common themes in the ways increasingly proficient people structure and use their knowledge in areas as diverse as chess, architecture, volleyball, shipboard navigation, and emergency room medicine (Ericsson, 1996). Identifiable kinds of things people do in certain kinds of situations are observed in domains and at levels of education quite different in their particulars. An example is the phenomenon of "design under constraint," which is clearly at the heart of engineering and architecture but is equally apropos in creative domains such as writing and everyday activities such as planning a vacation. Being able to recognize constraints, use strategies for dealing with them, and monitor how one is progressing are common to developing proficiency in any domain where one must design in the face of constraints. It is thus a schema that we, as assessment designers, want to recognize in any domain that is the target of assessment, and we want to be able to develop tasks that evince this aspect of proficiency in the context of the domain's particulars. This is the role that *design patterns* play.

Architects and software engineers use the term *design pattern* for knowledge structures that characterize recurring problems and approaches for dealing with them (Alexander, Ishikawa, & Silverstein,1977; Gamma, Helm, Johnson, & Vlissides, 1994). *Design patterns* organize experience across many particular situations in ways that help a designer recognize and tackle challenges such as planning work flow in a kitchen or generating software objects. *Design patterns* for assessment design likewise help domain experts and assessment specialists "fill in the slots" of an assessment argument built around recurring themes in learning (Mislevy et al., 2003). The PADI Project's *design patterns* focus on science inquiry and concern the transdisciplinary themes noted above such as inquiry cycles, knowledge representation, model-based reasoning (Hafner & Stewart, 1995), and scaffolded performance.

Table 2, adapted from Mislevy and Riconscente (2006), lists the attributes of a PADI *design pattern*, defines the attributes, and specifies which component of the assessment argument it represents. *Design patterns* are intentionally broad and non-technical, "centered around some aspect of KSAs, a design pattern is meant to offer a variety of approaches that can be used to get evidence about that knowledge or skill, organized in such a way as to lead toward the more technical work of designing particular tasks" (Mislevy & Riconscente, 2006, p. 72).

Table 2. Design Pattern Attributes, Definitions, and Corresponding Assessment Argument Components

Attribute	Definition	Assessment Argument Component
Rationale	Explain why this phenomenon is an important aspect of scientific inquiry	Warrant (underlying)
Focal Knowledge, Skills, and Abilities	The primary knowledge/skill/abilities targeted by this design pattern	Student Model
Additional Knowledge, Skills, and Abilities	Other knowledge/skills/abilities that may be required by this design pattern.	Student Model
Potential Observations	Some possible things one could see students doing that would give evidence about the knowledge/skills/abilities	Evidence Model
Potential Work Products	Modes, like a written product or a spoken answer, in which students might produce evidence about KSAs.	Task Model
Characteristic Features	Aspects of assessment situations that are likely to evoke the desired evidence.	Task Model
Variable Features	Aspects of assessment situations that can be varied in order to shift difficulty or focus.	Task Model

Table 3 presents an example of a *design pattern* developed during the PADI Project, entitled "Model Elaboration." A brief description of each attribute on the *design pattern* is presented. It is one of seven design patterns based on research on model-based reasoning, summarized in Table 4. Forty-three *design patterns* at varying levels of abstract and specificity have been developed in the project so far.

Table 3. "Model Elaboration" Design Pattern in PADI Design System

Attribute	Value(s)	Comments
Title	Model Elaboration	
Summary	This design pattern concerns working with mappings and extensions of given scientific models.	A central element of scientific inquiry is reasoning with models. This design pattern focuses on model elaboration, as a perspective on assessment in inquiry and problem-solving.
Rationale	Scientific models are abstracted schemas involving entities and relationships, meant to be useful across a range of particular circumstances. Correspondences can be established between them and real-world situations and other models. Students use and gain conceptual or procedural knowledge by working with an existing model.	Students' work is bound by the concept of an existing model (or models) so that their work includes an understanding of the constraints of the problem. Even though model elaboration does not involve the invention of new objects, processes, or states, it does entail sophisticated thinking and is an analogue of much scientific activity.
Focal Knowledge, Skills, and Abilities	▪ Establishing correspondence between real-world situation and entities in a given model ▪ Finding links between similar models (ones that share objects, processes, or states) ▪ Linking models to create a more encompassing model ▪ Within-model conceptual insights	This design pattern focuses on establishing correspondences among models and between models and real-world situations.
Additional Knowledge, Skills, and Abilities	▪ Familiarity with task (materials, protocols, expectations) ▪ Subject-area knowledge ▪ Reasoning within the model ▪ Model revision	According to the designer's purposes, tasks may stress or minimize demand for other KSAs, including content knowledge, familiarity with the task type, and other aspects of model-based reasoning, including reasoning within models and revising models.

Table 3. "Model Elaboration" Design Pattern in PADI Design System (Continued)

Attribute	Value(s)	Comments
Potential observations	▪ Qualities of mapping the corresponding elements between a real-world situation and a scientific model. ▪ Appropriateness of catenations of models across levels (e.g., individual-level and species-level models in transmission genetics) ▪ Correctness and/or completeness of explanation of modifications, in terms of data/model anomalies ▪ Identification of ways that a model does not match a situation (e.g., simplifying assumptions), and characterizations of the implications.	These are examples of aspects of things that students might say, do, or construct in situations that call for model elaboration. They are meant to stimulate thinking about the observable variables the designer might choose to define for assessment tasks addressing model elaboration.
Potential rubrics		
Characteristic features	Real-world situation and one or more models appropriate to the situation, for which details of correspondence need to be fleshed out. Addresses correspondence between situation and models, and models with one another.	Any task concerning model elaboration generated in accordance with this design pattern will indicate a model or class of models the student is to work with, and real-world situations and/or other models to which correspondences are to be established.
Variable features	▪ Is problem context familiar? ▪ Model provided or to be produced by student(s)? ▪ Experimental work or supporting research required? ▪ Single model or correspondence among models? ▪ How well do the models/data correspond?	

Table 4. Design Patterns for Model-Based Reasoning in Science

Model formation	Establishing a correspondence between some real-world phenomenon and a model, or abstracted structure, in terms of entities, relationships, processes, behaviors, etc. Includes scope and grain-size to model and determining which aspects of the situation(s) to address and which to leave out.
Model elaboration	Combining, extending, adding detail to a model, establishing correspondences across overlapping models. Often done by assembling smaller models into larger assemblages, or fleshing out more general models with more detailed models.
Model use	Reasoning through the structure of a model to make explanations, predictions, conjectures, etc.
Model articulation	Establishing mappings between qualitative entities and relationships in a model and their representation in an associated symbol system. Relevant in models with quantitative/symbolic components, as with the connections between conceptual and mathematical aspects of physics models.
Model evaluation	Assessing the correspondence between the model components and their real-world counterparts, with emphasis on anomalies and important features not accounted for in the model.
Model revision	Modifying or elaborating a model for a phenomenon in order to establish a better correspondence. Often initiated by model evaluation procedures.
Model-based inquiry	Working interactively between phenomena and models, using all of the aspects above. Emphasis on monitoring and taking actions with regard to model-based inferences vis a vis real-world feedback.

In the PADI design system, each *design pattern* is presented as an online form with "slots" for each attribute. When the *design pattern* is completed, it specifies elements that can be assembled into an assessment argument. The assessment designer, in collaboration with domain experts, teachers, and other key stakeholders in the assessment, would work together to complete the *design pattern* by filling in the slots. The Title and Summary attributes summarize the purpose and basic idea of the *design pattern*. The Rationale specifies the underlying warrant that links the target inferences and the kinds of tasks and evidence that support them.

Focal Knowledge, Skills, and Abilities (KSAs) come from the valued knowledge identified during the *domain analysis*. The KSAs specify the substance of the claim about students that the assessment tasks (built in accordance with the *design pattern*) will address. Additional KSAs are those Knowledge, Skills, and Abilities that may also be required to complete a task that is targeting the Focal KSAs, depending on design choices. For

example, certain familiarity with representational forms or mathematical operations may be presumed in an investigation that is meant to focus on experimental technique. Additional KSAs may be included in tasks intentionally, avoided, or dealt with by allowing student choice or accommodations. The point of including this attribute in the *design pattern* is to make task authors aware of design choices and their implications. In particular, Additional KSAs highlight possible explanations for poor performance that are based on knowledge or skills that the task demands other than the targeted, Focal KSA—sources of construct-irrelevant variance in Messick's (1989) terminology.

Potential Work Products are the student responses that provide information about the Focal KSAs. Potential Work Products are what students say, do, or make. Potential Observations are the particular aspects of the Work Products that constitute evidence. Potential Work Products are "nouns," whereas the Potential Observations are adjectives that describe qualities of the Work Products—such as "number of," "quality of," and "kind of." Potential Rubrics are verbs—evaluation techniques that can be used or adapted to identify and evaluate these qualities. In familiar terms, the Potential Rubrics are used to "score" Work Products and result in producing values for the Observations. All of these attributes concern ways of getting evidence about the targeted aspect of proficiency—and the wider the array, the better, so that assessment designers can see ways to obtain evidence in their particular situation in light of the various mix of costs, constraints, familiarity, and feedback to learning associated with different ways to get evidence.

Characteristic Features and Variable Features specify aspects of the situation in which students act and produce Work Products. Characteristic Features are those that all assessment tasks motivated by the *design pattern* should have in some form, as they are central to evoking evidence about the Focal KSAs. All tasks inspired by the "Design under Constraints" *design pattern* must involve a design goal, multiple constraints, and a medium for design. Variable Features address those aspects of the assessment that the assessment designer can implement in different ways—in some cases within specific constraints. Some the Variable Features in "Design under Constraints" are the design domain itself, the familiarity or novelty of the problem, whether the goal is explicit or implicit, whether the design problem is embedded in a larger task (for which other *design patterns* may prove helpful), the number, difficulty, and interactions of constraints, and whether the work is collaborative. In the "Building Scientific Explanations" *design pattern* the BioKIDS project created, the amount of scaffolding that a student receives is a key Variable Feature. Mislevy and Riconscente (2006) note that, "Within the constraints of the Characteristic Features, choosing different configurations of Variable Features allows a designer to provide evidence about the Focal KSAs but influence the level of difficulty, the degree of confounding with other knowledge, gather more or less evidence at lesser or greater costs, and so on" (p. 75).

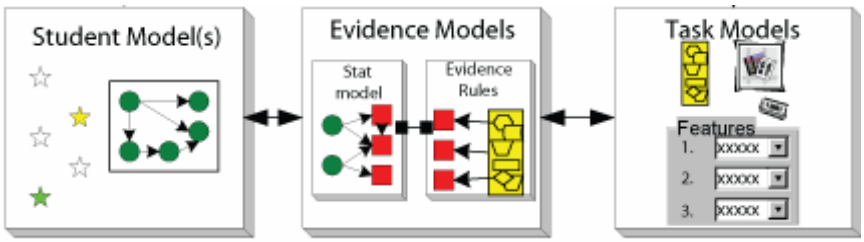
Work at the *domain modeling* layer is important for improving the practice of assessment, especially in light of the higher-level reasoning and capabilities for situated actions that research on learning brings to the fore. Experience in assessing these proficiencies is sparse and appears in contextualized exemplars—specific tasks are confounded with particular domains, psychological stances, knowledge representations, and delivery vehicles. Because proficiencies are their primary organizing category, *design patterns* focus attention

on the nature of proficiency one wants to assess. Delivery modes, response types, stimulus materials, and measurement models are secondary, determined in any particular setting to best instantiate the argument in light of the particular constraints and resources of that setting.

2.3 The Conceptual Assessment Framework

The *conceptual assessment framework* (CAF) concerns the technical specifications for the nuts and bolts of assessments—blueprints, as it were. The central models are the Student (or proficiency) Model, Evidence Models, and Task Models (Figure 4).[1] These models have their own internal logic and structures but are connected to each other through key elements described in the following paragraphs. An assessment argument laid out in narrative form at the *domain modeling* layer is now expressed in terms of designs for coordinated pieces of machinery such as measurement models, scoring methods, test assembly specifications, and requirements and protocols for delivery in the intended testing setting. In specifying the CAF, the assessment designer makes the decisions that will give shape to the assessment that will be generated. Details about task features, measurement models, stimulus material specifications, and the like are expressed in terms of knowledge representations that are tuned to constructing these elements and making sure they will operate coherently with one another. After the CAF has been specified, the assessment argument will have been expressed in concrete terms.

Figure 4. Graphic Summary of the Student, Evidence, and Task Models



There are considerable advantages to explicating the objects in this design layer. Having to construct coordinated forms helps organize the work of the different specialists involved in designing complex assessments. Because the models are themselves nearly independent, they are readily recombined when the same kinds of tasks are repurposed, from summative to formative uses, for example, by using a finer-grained Student Model with additional Observable Variables extracted from the same Work Products. Common data structures encourage the development of supported or automated processes for task creation (e.g., Irvine & Kyllonen, 2002), evaluating Work Products (e.g., Williamson, Mislevy, & Bejar, 2006), and assembling measurement models modularly (e.g., Rupp, 2002, von Davier, 2005). These features are especially important for computer-based tasks that are costly to author and implement, such as interactive simulations (see, for example, Neimi & Baker, 2005, and Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002 on task design; Luecht, 2002 on authoring and assembly; and Stevens & Casillas, 2006 on automated

[1] Defined abstractly in Mislevy, Steinberg, and Almond (2003), they can be implemented in different specific forms, as in Wilson's (2005) four-model parsing of the system and PADI template objects that catenate evidence and task models.

scoring). Bringing down the costs of such tasks requires exploiting every opportunity to reuse arguments, structures, processes, and materials.

The Student Model addresses what the assessment designer is trying to measure as expressed in terms of one or more variables that reflect aspects of students' proficiencies. A probability distribution over these variables expresses what is known about their values for a given student, at a given point in time, after a given set of observations has been obtained. The aspects of proficiency should be consistent with the conception of knowledge in the targeted domain, but design choices must still be made as to the number, character, and granularity of Student Model Variables so that they best serve the purpose of the assessment. A Student Model could contain a single variable in order to characterize students simply in terms of an overall proficiency in a domain of tasks. When several aspects of proficiency are involved, students vary in their profiles, and tasks require different mixes of the proficiencies, a vector-valued Student Model and a multivariate probability distribution may be used to express what is known about a student. Such models are suited to providing feedback in greater detail and to sorting out patterns of proficiency from complex performances such as investigations or interactive language tasks. Establishing a model as mechanically distinct from evidence about it allows for flexibility in assessing proficiencies with different kinds of tasks in different contexts.

The Task Model addresses the environment in which the test takers will say, do, or make something to provide the data about what they know or can do. A key decision is specifying the form in which students' performances will be captured, i.e., the Work Product(s)—for example, a choice among alternatives, an essay, a graph, a formula, a painting, a sequence of steps in an investigation, or the locations of icons dragged into a diagram. In computer-based testing with complex tasks, reusing the same underlying Work Product forms can streamline authoring, implementation, and evaluation (Scalise, 2003). The assessment designer also specifies the forms and the salient features of materials that will be necessary as directives, manipulatives, stimulus materials, as well as what features of the environment must be present to administer the assessment as intended. For example, what resources must be available to the test taker or what degree of scaffolding can be provided by the teacher? These decisions are guided by discussions in *domain modeling*, in terms of Characteristic and Variable Task Features. Again, efficiencies accrue from the reuse of structures, processes, activity flows, tools, and materials. Of particular importance are schemas for tasks, suggested in *domain modeling* and now implemented that aid item writers—not a new idea, by any means. Witness, for example, Bormuth's (1970) algorithm for generating comprehension tasks as one whose time has arrived in light of information-processing methods for representing information in learning domains (e.g., Marshall, 1995).

How does student performance, captured in the form of Work Products, update beliefs about a student? The Evidence Model bridges the Student Model Variables and the Task Model. There are two components to the Evidence Model—the evaluation component and the Measurement Model. The evaluation component indicates which aspects of the work are important and how they will be evaluated. These aspects of the work are referred to as Observable Variables, or "item scores" in a more familiar but more constrained usage. Explicit Evaluation Procedures are specified to indicate how the values of Observable

Variables are determined from the students' Work Products, be they algorithms for automated scoring procedures or rubrics, examples, and training materials for humans. Again to promote reuse and modular construction of assessments, different Evaluation Procedures can be used to extract different Observable Variables from the same Work Products when tasks are used for different purposes, and different implementations can be used to extract the same Observable Variables from the same Work Products, as when different venders use different algorithms to score tasks or both human judges and automated scoring of essays produces ratings in the same form.

> The data that are generated in the evaluation component are synthesized across tasks in the Measurement Model component—specifically, in educational and psychological Measurement Models, conditional distributions of Observable Variables given student model variables. The modular and tailored construction of Measurement Models mentioned above assembles pieces in the form of IRT, latent class, Bayes net, or other model fragments, in accordance with the nature of the Student Model and Observable Variables. Two related developments of particular interest are the assembly of Measurement Models in accordance with the values of Task Model Variables (which in turn reflect the theory underlying task construction) and incorporating Task Model Variables into Measurement Models, reducing or eliminating the need to calibrate items from empirical data alone (Embretson, 1998). Again, much can be gained when the evidentiary relationships in complex tasks and multivariate Student Models are captured in assemblies of possibly complex Measurement Model fragments, structures that can be used with many tasks with different particulars as to content and materials (Mislevy et al., 2002). The vector observations of the BioKIDS conjecture-and-explanation tasks use the same Observable Variable definitions and bundled measurement structure that accounts for their conditional dependence (Gotwals & Songer, 2006). In this way, task authors can create unique complex tasks from *template* components and know ahead of time "how to score them."

In the PADI assessment design system, the specification of the CAF takes place in the form of completing PADI *task templates*. Figure 5 is a high-level representation of a PADI *template* as a Unified Modeling Language (UML) diagram, and Figure 6-Figure 8 show selected objects from one of the *templates* from the BioKIDS project using the design-system interface. Figure 7 is the same Measurement Model object as in Figure 8 but seen as the XML file the designer creates when using the interface. This form is better suited to sharing, transporting, and serving as input or output to automated procedures in implementation or delivery. The PADI *design patterns* and *templates* can be considered what Collins & Ferguson (1993) call "epistemic forms"—structures that embody key principles in a domain, and the act of filling them out creates knowledge (see Brecht, Mislevy, Haertel, & Haynie, 2005, for a discussion of PADI in terms of epistemic forms and games). PADI *task templates* guide the creation of families of tasks to be specified with details of materials and task settings in the *assessment implementation* layer.

Figure 5. High-level UML Representation of the PADI Object Model

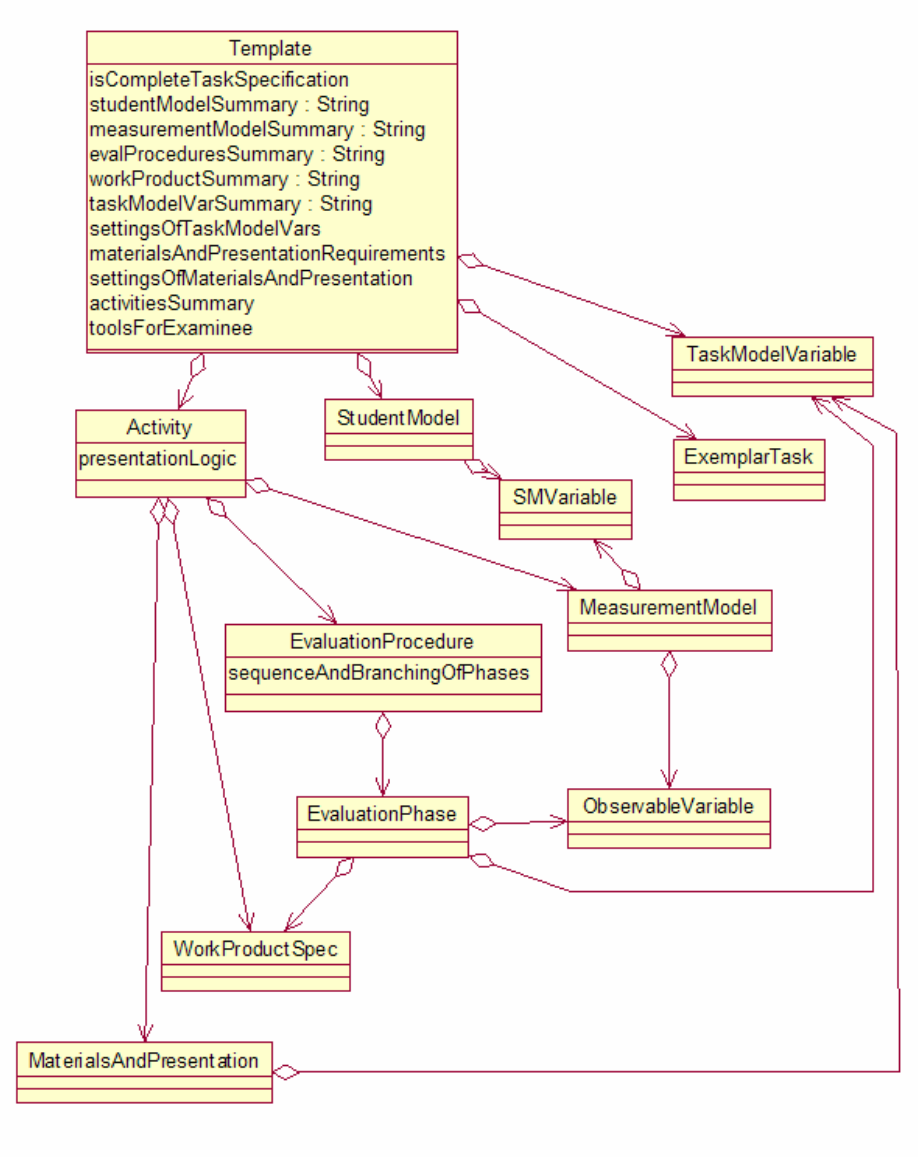


Figure 6. A BioKIDS Template within PADI Design System

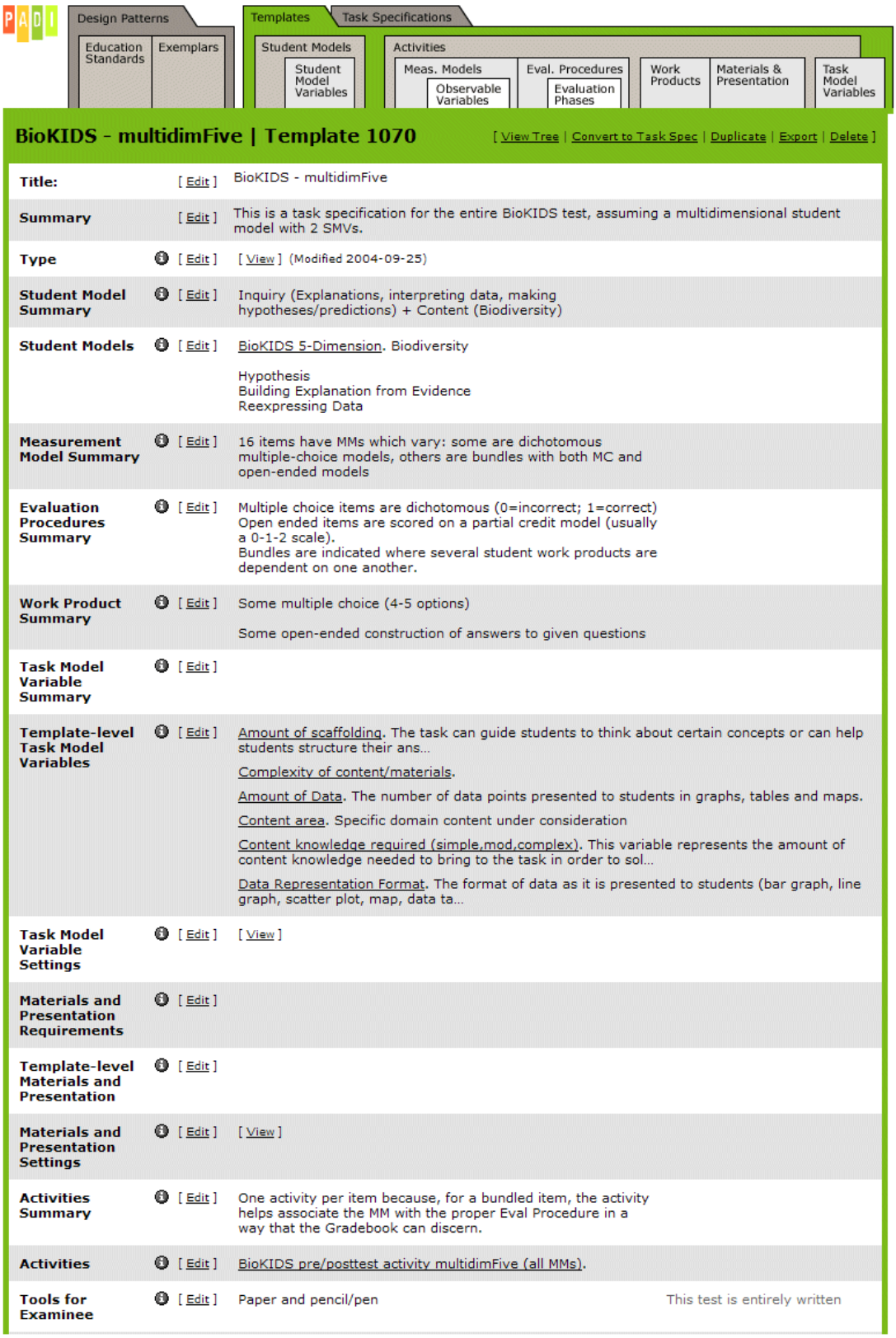


Figure 7. A BioKIDS Measurement Model within the PADI Design System, as Viewed through the User Interface

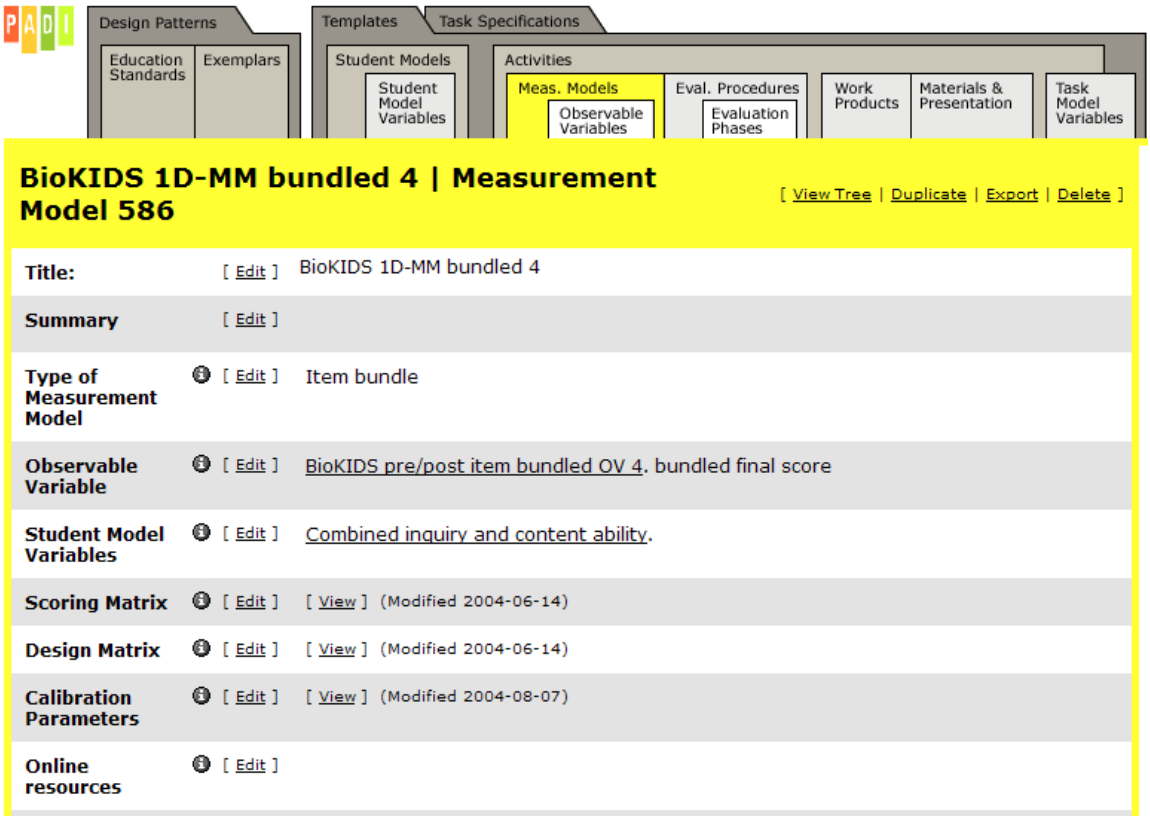


Figure 8. The XML Representation of a Measurement Model within the PADI Design System

```
− <MEASUREMENT_MODEL_TYPE NODE_TITLE="BioKIDS 1D-MM bundled 4"
  NODE_TYPE_VERSION="4.80" ident="586">
    <TYPE_OF_MEAS_MODEL PART_LABEL="Type of Measurement Model" ATTRIBUTE_ID="2651"
    ATTRIBUTE_VALUE="52" VALUE_IN_PICKLIST="Item bundle"/>
  − <RELATED PART_LABEL="Observable Variable" PART_TYPE="OBSERVABLE_VARIABLE"
    RELATION_TYPE_NAME="DEST_IS_PART_OF_SRC" PART_NUM="1">
    − <OBSERVABLE_VARIABLE NODE_TITLE="BioKIDS pre/post item bundled OV 4"
      NODE_TYPE_VERSION="1.60" ident="587">
        <NODE_ANNOTATION>bundled final score</NODE_ANNOTATION>
        <OV_CATEGORY PART_LABEL="Categories (possible values)" ATTRIBUTE_ID="2656"
        ATTRIBUTE_COMMENT="claim 0, evidence 0" ATTRIBUTE_ORDER="1" ATTRIBUTE_VALUE="0"/>
        <OV_CATEGORY PART_LABEL="Categories (possible values)" ATTRIBUTE_ID="2655"
        ATTRIBUTE_COMMENT="claim 0, evidence 1" ATTRIBUTE_ORDER="2" ATTRIBUTE_VALUE="1"/>
        <OV_CATEGORY PART_LABEL="Categories (possible values)" ATTRIBUTE_ID="2652"
        ATTRIBUTE_COMMENT="claim 1" ATTRIBUTE_ORDER="3" ATTRIBUTE_VALUE="2"/>
        <OV_CATEGORY PART_LABEL="Categories (possible values)" ATTRIBUTE_ID="2654"
        ATTRIBUTE_COMMENT="claim 1, evidence 1" ATTRIBUTE_ORDER="4" ATTRIBUTE_VALUE="3"/>
        <OV_CATEGORY PART_LABEL="Categories (possible values)" ATTRIBUTE_ID="2653"
        ATTRIBUTE_COMMENT="claim 1, evidence2" ATTRIBUTE_ORDER="5" ATTRIBUTE_VALUE="4"/>
      </OBSERVABLE_VARIABLE>
    </RELATED>
  − <RELATED PART_LABEL="Student Model Variables" PART_TYPE="STUDENT_MODEL_VARIABLE_TYPE"
    RELATION_TYPE_NAME="DEST_IS_PART_OF_SRC" PART_NUM="2">
    − <STUDENT_MODEL_VARIABLE_TYPE NODE_TITLE="Combined inquiry and content ability"
```

2.4 Assessment Implementation

The *assessment implementation* layer of ECD concerns constructing and preparing all of the operational elements specified in the CAF. This includes authoring tasks, producing test forms, finalizing rubrics or automated scoring rules, and determining parameters for conditional probabilities in Measurement Model fragments. All of these activities are familiar in current tests and are often quite efficient. What the ECD approach provides is the coordinated rationale for each, all the way back through the assessment argument, and the use of structures which at every stage highlight opportunities for reuse and interoperability. These capabilities leverage the value of systems for authoring or generating tasks, calibrating items, presenting materials, and interacting with examinees. For an example, see the work of Baker and her colleagues in the IERI project "Assessments to Support the Transition to Complex Learning in Science" (Baker,2002; Neimi, 2005; and Vendlinsky, Neimi, & Baker, in press).

Developing tools for this layer has not been the focus of PADI. However, the BEAR Assessment Center at the University of California, Berkeley has produced a Scoring Engine and a Calibration Engine as part of the PADI Project. "Wizards" have been developed to help test developers create individual tasks from PADI *templates* in the GLOBE exemplar, and to help psychometricians assemble complex Measurement Models. In addition, a data management tool referred to as Gradebook, developed at the University of Maryland, is also available to support passing data from the design system to the Scoring and Calibration Engines in formats that are compatible with the IMS/QTI protocol. The reader is referred to the PADI technical report series for details (http://padi.sri.com/publications.html), as some of this work is complete and documented while the rest still is being documented.

2.5 Assessment Delivery

The *assessment delivery* layer is where students interact with tasks, their performances are evaluated, and feedback and reports are produced. The PADI project uses the *four-process delivery* system described in Almond, Steinberg, and Mislevy (2002), which is also the conceptual model underlying the IMS/QTI specifications. This parsing of activities can be used to describe computer-based testing procedures, but with appropriate interpretations, paper-and-pencil tests, informal classroom tests, or tutoring systems. Common language, common data structures, and a common partitioning of activities again promote reuse of objects and processes and interoperability across projects and actors. The processes pass messages to one another in a pattern determined by the test's purpose. All of the messages are either data objects specified in the CAF (e.g., parameters, stimulus materials) or are produced by the student or other processes in data structures that are specified in the CAF (e.g., Work Products, values of Observables Variables).

Assessment operation is represented as four principal processes. The *activity selection process* selects a task or activity from the task library or creates one in real time in accordance with *templates* that are instantiated in light of what is known about the student or the situation. The *presentation process* is responsible for presenting the task to the student, managing the interaction, and capturing Work Products. Work Products are then passed to the *evidence identification process*, or task-level scoring, which performs

item-level response processing according to the methods specified in the evaluation rules in the Evidence Model. Values of the Observable Variables are sent to the *evidence accumulation process,* or test-level scoring, which summarizes evidence in terms of probability distributions for the Student Model Variables via the Measurement Model. In adaptive tests the evidence accumulation process provides information to the *activity selection process* to help determine what to do next. The *four-process delivery architecture* is compliant with Question and Test Interoperability (QTI) standards to help assessment developers share materials and processes across assessment systems and platforms.

As with *assessment implementation*, many *assessment delivery* systems exist and many are quite efficient in the settings for which they were developed. Reusability and interoperability are the watchwords here, particularly for Web- and computer-based testing. This, of course, is the idea behind IMS/QTI standards for electronic assessment. The ECD framework facilitates the development of assessments and assessment materials and processes that accord with current specifications and with the overarching principles more generally. Such efforts help bring down the costs of developing and delivering innovative assessments at the large scale required in statewide testing.

PADI also has not focused on *assessment delivery*. The PADI Mystery Powders example (Siebert, Hamel, Haynie, & Mislevy, 2006), however, does illustrate the connections among the CAF objects, implementation, and delivery in a computer simulation of a well known hands-on investigation used in large-scale assessments.

3.0 Conclusion: Aren't These Just New Words for Things We Already Do?

So what is the bottom line? Is evidence-centered design just a bunch of new words for things we already are doing? There is a case to be made that it is. All of the innovations sketched above—in cognitive psychology, learning in domains, measurement models, task design, scoring methods, Web-based delivery, and more—have been developed by thousands of researchers across many fields of study without particular regard for ECD. So too have new assessment types arisen, each in their stead. And established and efficient procedures for familiar assessments have been evolving for decades, continually improving in increments. Changing vocabularies and representational forms would like as not slow them down, as long as their current goals and processes suit their aims and resources.

But efficiency just within assessments can impede efficiency across assessments. Efficient work within large-scale assessments takes place because each contributor knows his or her job, but connections among the work they do remain implicit. Modifying an assessment is difficult because what appear to be improvements from one vantage point conflict with other parts of the system in unforeseen ways. Elements or processes that could in principle be shared across assessments are not, because their data structures are incompatible or delivery stages are collapsed differently. Analyzing existing assessments in terms of common vocabularies and representational forms across the ECD layers helps bring out the fundamental similarities across assessments that can look very different on the surface and alert us to opportunities for improvement.

Even greater gains accrue for new kinds of tests, both conceptually and technically. The conceptual advantages come from grounding the design process from the beginning on the assessment argument in the form of tools like *design patterns*. Thinking through how to assess new or complex proficiencies as in science inquiry and task-based language assessment is best done at a layer that focuses on the conceptual argument and is capable of being implemented in different ways rather than being entangled with implementation or delivery choices. This work is a natural bridge between conceptual developments reflected in research and standards, on the one hand, and practical testing methods on the other. Work at this layer improves practice in its own way for large-scale, classroom, certification, or other testing venues.

The technical advantages come about because no existing process can be pulled off the shelf and implemented in toto. More original design work is therefore necessary to rationalize, implement, and deliver a new kind of, say, simulation task. ECD's language, representational forms, and unified perspective of the assessment enterprise guide planning and coordinate work in developing tasks and operational methods. They entail laying out the assessment argument, clarifying design choices, and coordinating the development of operational elements. At every step along the way, they encourage the recognition and exploitation of efficiencies from reuse and compatibility. Moreover, they provide a principled framework to work through accommodation and universal design at the level of the validity argument as well as delivery issues (Hansen, Mislevy, Steinberg, Lee, & Forer, 2005).

Evidence-centered design is a framework, then, that does indeed provide new words for things we already are doing. That said, it helps us understand what we are doing at a more fundamental level. And it sets the stage for doing what we do now more efficiently and learning more quickly how to assess in ways that we do not do now, either because we don't know how or can't afford to.

References

Adams, R., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A pattern language: Towns, buildings, construction.* New York: Oxford University Press.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, *1*(5). Retrieved 6/25/2006 from http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

American Association for the Advancement of Science (AAAS, 1993). *Benchmarks for Scientific Literacy*. New York: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Baker, E. L. (1997, Autumn). Model-based performance assessment. *Theory Into Practice, 36,* 247-254.

Baker, E. L. (2002). Design of automated authoring systems for tests. *Proceedings of Technology and assessment: Thinking ahead proceedings from a workshop*." (pp. 79-89). Published of Collection: National Research Council, Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education.

Baxter, G. & Mislevy, R. J. (2004). *The case for an integrated design framework for assessing science inquiry* (CSE Technical Report 638). Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global E-learning program. *The International Journal of Testing, 4*, 295-301.

Bormuth, J. R. (1970). *On the theory of achievement test items.* Chicago: University of Chicago Press.

Brand, S. (1994). *How buildings learn: What happens after they're built*. New York: Viking-Penguin.

Brecht, J., Mislevy, R., Haertel, G. D., & Haynie, K. C. (2005, April). *The PADI design system as a complex of epistemic forms and games.* Paper presented annual meeting of the American Educational Research Association, Montreal.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Research, 18*(1), 32-42.

Chung, G. K., Delacruz, W. K., Dionne, G. B., & Bewley, W. L. (2003, December). Linking assessment and instruction using ontologies. Proceedings of the I/ITSEC, Orlando, FL.

Cisco Systems (2000). *Internetworking technology basics* (3rd ed.). Indianapolis, IN: Author.

Clyman, S. G., Melnick, D. E., & Clauser, B. E. (1999). Computer-based case simulations from medicine: Assessing skills in patient management. In A. Tekian, C. H. McGuire, W. C. McGahie (Eds.), *Innovative simulations for assessing professional competence* (pp.29-41). Chicago: University of Illinois, Department of Medical Education.

Collins, A., & Ferguson, W. (1993). Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational Psychologist*, *28*(1), 25-42.

Embretson, S. E. (1985). A general latent trait model for response processes. *Psychometrika, 49,* 175-186.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Ericsson, K. A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed), *The Road to Excellence: The Acquisition of Expert Performances, Sports, and Games.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns.* Reading, MA: Addison-Wesley.

Gotwals, A. W., & Songer, N. B. (2006). *Cognitive predictions: BioKIDS implementation of the PADI assessment system* (PADI Technical Report 10). Menlo Park, CA: SRI International.

Greeno, J. G. (1983). Conceptual entities. In D. Gentner & A. L. Stevens (Eds.), *Mental models.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hafner, R., & Stewart, J. (1995). Revising explanatory models to accommodate anomalous genetic phenomena: Problem solving in the "context of discovery." *Science Education*, *79*, 111-146.

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests within a validity framework. *System: An International Journal of Educational Technology and Applied Linguistics, 33,* 107-133.

Huber, A. E., Songer, N. B., and Lee, S.-Y. (2003). *A curricular approach to teaching biodiversity through inquiry in technology-rich environments*. Paper presented at the annual meeting of the National Association of Research in Science Teaching (NARST), Philadelphia.

IMS Global Learning Consortium (2000). *IMS question and test interoperability specification: A review* (White Paper IMSWP-1 Version A). Burlington, MA: Author.

Irvine, S. H., & Kyllonen, P. C. (Eds.), (2002). *Item generation for test development*. Hillsdale, NJ: Erlbaum.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527-535.

Lehrer, R., & Schauble, L. (2002). Symbolic communication in mathematics and science: Co-constituting inscription and thought. In E. D. Amsel & J. Byrnes (Eds.), *Language, literacy, and cognitive development. The development and consequences of symbolic communication*. (pp. 167-192). Mahwah, NJ: Lawrence Erlbaum Associates.

Luecht, R. M. (2002). *From design to delivery: Engineering the mass production of complex performance assessments.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Markman, A. B. (1999). *Knowledge representation.* Mahwah, NJ: Erlbaum.

Marshall, S. P. (1995). *Schemas in problem solving.* Cambridge: Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90). Mahwah, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R.G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15,* 363-378.

National Research Council (1996). *National Science Education Standards.* Washington: National Academy Press.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment, J. Pellegrino, R. Glaser, & N. Chudowsky (Eds.). Washington DC: National Academy Press.

Niemi, D. (2005, April). Assessment objects for domain-independent and domain specific assessment. In F. C. Sloane & J. W. Pellegrino (co-Chairs), *Moving technology up-design requirements for valid, effective classroom and large-scale assessment*. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Niemi, D., & Baker, E. L. (2005, April). Reconceiving assessment shortfalls: System requirements needed to produce learning. In F. C. Sloane & J. W. Pellegrino (co-Chairs), *Moving technology up-design requirements for valid, effective classroom and large-scale assessment*. Presentation at the annual meeting of the American Educational Research Association, Montreal.

Raymond, M., & Neustel, S. (2006). Determining test content of credentialing examinations. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 181-224). Mahwah, NJ: Erlbaum

Rupp, A. A. (2002). Feature selection for choosing and assembling measurement models: a building-block-based organization. *International Journal of Testing, 2,* 311-360.

Scalise, K. (2003). *Innovative item types and outcome spaces in computer-adaptive assessment: A literature survey*. Berkeley Evaluation and Assessment Research (BEAR) Center, University of California at Berkeley.

Shute, V. J. & Torreano, L., & Willis, R. (2000). DNA: Towards an automated knowledge elicitation and organization tool. In S. P. Lajoie (Ed.) *Computers as Cognitive Tools, Volume 2.* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 309-335.

Siebert, G., Hamel, L., Haynie, K., Mislevy, R., & Bao, H. (2006). *Mystery Powders: An application of the PADI design system using the four-process delivery system*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Simon, H.A. (2001). *The sciences of the artificial* (4th ed.). Cambridge, MA: MIT Press.

Songer, N.B. (2004) *Evidence of complex reasoning in technology and science: Notes from Inner City Detroit, Michigan, USA.* IPSI-2004 Pescara Conference, Italy.

Stevens, R., & Casillas, A. (2006). Artificial neural networks. In D.M. Williamson, R.J. Mislevy, & I.I. Bejar (Eds.), *Automated scoring of complex tasks in computer based testing* (pp. xxx-xxx). Mahwah, NJ: Erlbaum Associates.

Stewart, J., & Hafner, R. (1994). Research on problem solving: Genetics. In D. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp 284-300). New York: Macmillan.

Toulmin, S.E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions*. Cambridge, MA: MIT Press.

Vendlinski, T. P., Niemi, D., & Baker, E. L. (in press). Objects and templates in authoring problem-solving assessments. In E. L. Baker, J. Dickieson, W. Wulfeck, & H. F. O Neil (Eds.), *Assessment of problem solving using simulations*. Mahwah, NJ: Erlbaum.

Von Davier, M. (2005). A class of models for cognitive diagnosis. *Research Report RR-05-17*. Princeton, NJ: ETS.

Webb, N. (2006). Identifying content for student achievement tests. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 155-180). Mahwah, NJ: Erlbaum.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16,* 3-118.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.) (2006). *Automated scoring of complex tasks in computer based testing*. Mahwah, NJ: Erlbaum Associates.

Wilson, M. R. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Sponsor
The National Science Foundation, Grant REC-0129331

Prime Grantee
SRI International. *Center for Technology in Learning*

Subgrantees
University of Maryland
University of California, Berkeley. *Berkeley Evaluation & Assessment Research (BEAR) Center* and *The Full Option Science System (FOSS)*
University of Michigan. *BioKIDs*

SRI International